Published online 31 May 2012 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* ***W249–W256***
doi:10.1093/nar/gks481

DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry

Yao Chi Chen[1], Jon D. Wright[1,2] and Carmay Lim[1,3,*]

[1]Institute of Biomedical Sciences, [2]Genomics Research Center, Academia Sinica, Taipei 115, and [3]Department of Chemistry, National Tsing Hua University, Hsinchu 300, Taiwan

Received February 6, 2012; Revised April 27, 2012; Accepted May 4, 2012

ABSTRACT

DR_bind is a web server that automatically predicts DNA-binding residues, given the respective protein structure based on (i) electrostatics, (ii) evolution and (iii) geometry. In contrast to machine-learning methods, DR_bind does not require a training data set or any parameters. It predicts DNA-binding residues by detecting a cluster of conserved, solvent-accessible residues that are electrostatically stabilized upon mutation to Asp^-/Glu^-. The server requires as input the DNA-binding protein structure in PDB format and outputs a downloadable text file of the predicted DNA-binding residues, a 3D visualization of the predicted residues highlighted in the given protein structure, and a downloadable PyMol script for visualization of the results. Calibration on 83 and 55 non-redundant DNA-bound and DNA-free protein structures yielded a DNA-binding residue prediction accuracy/precision of 90/47% and 88/42%, respectively. Since DR_bind does not require any training using protein–DNA complex structures, it may predict DNA-binding residues in novel structures of DNA-binding proteins resulting from structural genomics projects with no conservation data. The DR_bind server is freely available with no login requirement at http://dnasite.limlab.ibms.sinica.edu.tw.

INTRODUCTION

Interactions between proteins and DNA play essential roles for life. For example, protein–DNA interactions control gene regulation, cell replication and transcription, as well as DNA repair. Furthermore, many of these DNA-binding proteins are involved in human diseases such as neurological disorders, e.g. TDP-43 (1), and cancer; e.g. p53 (2). Consequently, identifying the key amino acid residues involved in DNA recognition is critical for understanding these important biological processes. It also guides which residues to mutate in experimental studies.

Several methods and web servers have been developed to predict DNA-binding residues from the protein 1D sequence or 3D structure. Methods that predict DNA-binding residues using only the protein sequence generally employ machine-learning algorithms such as a neural network (3–5), a Naïve Bayes classifier (6), a support vector machine (7–12), random forest (13,14), or decision trees (C4.5 algorithm) (15). These algorithms usually employ amino acid physicochemical properties, sequence conservation, the local sequence context, solvent accessibility and/or secondary structure. Publicly available web servers that implement sequence-based methods for predicting DNA-binding residues include DBS-PRED (3), DBS-PSSM (5), DNABindR (6), DP-Bind (8), DISIS (9), BindN-rf (14), BindN+ (12), NAPS (15) and MetaDBSite (16). Methods that use the protein structure, if available, generally improve the DNA-binding site prediction, as they replace the predicted solvent accessibility, hydrophobicity and secondary structure in sequence-based methods with observed ones and can additionally employ energies or frequencies, computed from the atomic coordinates, as well as experimental geometrical features. Structure-based methods for predicting DNA-binding residues employ mostly electrostatic potentials in conjunction with other features such as surface/solvent accessibility, the protein surface shape, amino acid conservation, propensity, hydrophobicity and hydrogen-bonding potential and structural motifs (17–22), or high-frequency residue fluctuations (23). Servers that

*To whom correspondence should be addressed. Tel: +886 2 2652 3031; Fax: +886 2 2788 7641; Email: carmay@gate.sinica.edu.tw

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

implement structure-based methods for predicting DNA-binding residues include PreDs (24), DISPLAR (25), DBD-Hunter (26) and DNABINDPROT (23).

In our previous work (27), we had developed a structure-based DNA-binding residue prediction method based on (i) electrostatics, (ii) conservation and (iii) geometry with the following rationale: (i) DNA-binding residues contain electropositive atoms, which would be in an unfavorable electrostatic environment in the absence of DNA or water; thus replacing one of these residues with a negatively charged Asp^-/Glu^- would alleviate the electrostatic repulsion among the electropositive atoms in the gas phase; (ii) DNA-binding residues and residues in the vicinity, which form a cluster of spatially interacting residues, are usually highly conserved within the same family due to their critical functional roles and (iii) DNA-binding residues have been observed to be located on surface patches, as opposed to clefts/cavities for RNA-binding residues and enzyme substrates. In this work, we have implemented our DNA-residue prediction method for public use in a web server, DR_bind (http://dnasite.limlab.ibms.sinica.edu.tw). Whereas our published method for predicting DNA-binding sites had been tested on a non-redundant set of 56 DNA-bound and 23 DNA-free non-homologous protein structures (27), DR_bind was tested herein using an updated non-redundant set of 83 DNA-bound and 55 DNA-free structures (referred to as Data sets I and II, respectively). DR_bind was also tested using a protein–DNA docking benchmark containing 47 unbound–bound structures (28) and 15 non-redundant DNA-bound protein structures with no or insufficient homologous seqeunces to compute conservation scores reliably. In contrast to current DNA-binding residue prediction servers, DR_bind is based on physical principles of binding thermodynamics (29) and does not require training on a set of protein–DNA complexes or any parameters. Hence, DR_bind would be an opportune addition since structures of DNA-binding proteins have been rapidly rising.

METHODS

Data sets used

DR_bind was tested using four data sets: I—83 non-redundant DNA-bound protein structures, II—55 non-redundant DNA-free protein structures, III—47 bound–unbound structures from the protein–DNA benchmark version 1.2 (28) and IV—15 non-redundant DNA-bound protein structures with no, or insufficient homologs to compute conservation profiles reliably. To create Data set I, all available X-ray structures of DNA-bound proteins solved to $\leq$3-Å resolution were obtained from the current Protein Data Bank (PDB) (30). These protein chains were grouped according to their Class, Architecture, Topology and Homologous superfamily (CATH) codes (31). For each group of protein structures with the same CATH code, the structure with the best resolution was selected as the representative one. If any of these representative proteins share >30% sequence identity, the protein with the longer sequence was kept, while the others were discarded. This yielded 83 DNA-bound proteins that are sequentially and structurally non-homologous with conservation data (Supplementary Table S1), whereas the remaining 12 proteins had no conservation profiles from the ConSurf-DB database (http://consurfdb.tau.ac.il/) (32).

Data set II was derived from Data set I by searching each of the 83 DNA-bound proteins with conservation data for highly homologous proteins (sharing $\geq$90% sequence identity) with DNA-free structure(s) using the SAS tool (http://www.ebi.ac.uk/thornton-srv/databases/sas/); if multiple DNA-free structures were found, the structure that showed the largest root-mean-square deviation (RMSD) from the DNA-bound structure using the SSAP program (33) was chosen as the representative one. This yielded 55 bound–unbound structures with a wide range of RMSDs (0.3–33 Å). The PDB entries of the DNA-bound and free protein structures, the sequence identity between the DNA-bound and the respective free proteins computed using global alignment with ClustalW1.83 (34) and their RMSD values are given in Supplementary Table S1.

Data set III is a protein–DNA docking benchmark containing 47 bound–unbound structures, of which 13 were classified as 'easy', 22 as 'intermediate' and 12 as 'difficult' cases for docking depending on the interface RMSD values between the DNA-bound and corresponding free structures. 'Easy', 'intermediate' and 'difficult' structures were defined by interface RMSD values ranging from 0 to 2 Å, 2 to 5 Å, >5 Å, respectively. Data set III differs from Data set II in that it includes: (i) protein structures deposited in the September 2007 RCSB PDB; (ii) structurally homologous proteins with the same CATH code; (iii) free NMR structures; and (iv) 15 structures without conservation data from ConSurf-DB.

To create Data set IV, the 12 proteins excluded from Data set I and the 15 proteins from the benchmark set, which lack conservation profiles from ConSurf-DB, were grouped according to their CATH codes. For each group of protein structures with the same CATH code, the best resolution structure was selected as the representative one. This yielded 15 non-redundant proteins sharing <30% pairwise sequence identity (Supplementary Table S2).

Definitions

A residue was considered to bind DNA if it contains one or more non-hydrogen atom within van der Waals contact or hydrogen-bonding distance to the non-hydrogen atom of its binding partner directly or indirectly via a bridging water molecule. HBPLUS (35) was used to compute all possible hydrogen bonds and van der Waals contacts, which are defined by a donor atom to an acceptor atom distance $\leq$3.5 and $\leq$4.0 Å, respectively. An amino acid X is considered accessible for interacting with DNA if the percent ratio of its side chain solvent-accessible surface area in the protein to that in the tripeptide, –Gly–X–Gly–, is >5% (17,36). MOLMOL (37) was used to compute the relative solvent-accessible surface area of each amino acid from the protein structure using a solvent probe radius of 1.4 Å.

Geometry

Since DNA-binding sites are found on a protein surface, surface patches were generated by defining the C^{α} atom of each residue as an origin of a patch and including all residues whose C^{α} atoms were within 10 Å of the origin in the patch. Non-identical patches with more than five solvent-accessible residues were used in computing the average electrostatic energy change and conservation (see below).

Electrostatics

Given a *l*-residue DNA-binding protein structure, all Asp/Glu residues were deprotonated, while Arg/Lys residues were protonated; His residues were protonated or deprotonated depending on the availability of hydrogen bond acceptors in the structure. Next, *l* mutant structures were generated by replacing Ala, Asn, Asp, Cys, Gly, Ser, Thr or Val in the wild-type structure to Asp^{-} and the other residues to Glu^{-}. The side chain replacements were carried out using SCWRL (38), followed by energy minimization with heavy constraints on all heavy atoms using AMBER (39) to relieve any bad contacts. Based on the wild-type/mutant structures, the gas-phase ($\varepsilon = 1$) electrostatic energy of the wild-type ($E^{\text{elec}}{}_{\text{wt}}$) or mutant ($E^{\text{elec}}{}_{\text{mut}}$) protein in the 'folded' state relative to that in an 'extended reference' state ($E'^{\text{elec}}{}_{\text{wt}}$ or $E'^{\text{elec}}{}_{\text{mut}}$) was computed using AMBER (39) with the all-hydrogen-atom AMBER force field (40). In this extended reference state, the residues do not interact with one another; hence, the electrostatic energy difference between the wild-type ($E'^{\text{elec}}{}_{\text{wt}}$) or mutant ($E'^{\text{elec}}{}_{\text{mut}}$) 'unfolded' protein is equal to the difference between the electrostatic energies of the native residue at position *i* ($E'^{\text{elec}}{}_{i}$) and the corresponding mutant Asp^{-}/Glu^{-} ($E'^{\text{elec}}{}_{\text{D/E}}$). The change in the gas-phase electrostatic energy $\Delta\Delta E^{\text{elec}}{}_{i}$, upon mutation of residue *i* to Asp^{-}/Glu^{-} is given by:

$$\Delta\Delta E_i^{\text{elec}} = (E_{\text{mut},i}^{\text{elec}} - E_{\text{wt}}^{\text{elec}}) - (E_{\text{D/E}}'^{\text{elec}} - E_i'^{\text{elec}}) \quad (1)$$

The average electrostatic energy change $<\Delta\Delta E^{\text{elec}}>_i$ of the $N^{\text{aa}}{}_i$ residues comprising surface patch *i* was computed from:

$$< \Delta\Delta E^{\text{elec}} >_i = \sum \Delta\Delta E_j^{\text{elec}} / N_i^{\text{aa}} \quad (2)$$

where the summation in Equation (2) is over all residues in patch *i*.

Conservation

For a given DNA-binding protein, the conservation score C_i of residue *i* was obtained from the ConSurf-DB database (32) or ConSurf server (41–43). The C_i score is an integer number, ranging from 1 (for a rapidly evolving, highly variable residue) to 9 (for a slowly evolving, conserved residue). The average conservation $<C>_i$ of the $N^{\text{aa}}{}_i$ residues comprising surface patch *i* was computed from:

$$< C >_i = \sum C_j / N_i^{\text{aa}} \quad (3)$$

DNA-binding residue prediction

To determine the DNA-binding residues in a given protein, the distinct patches were ranked according to the $<\Delta\Delta E^{\text{elec}}>_i$ values so that the top-ranked cluster had the most favorable (most negative) $<\Delta\Delta E^{\text{elec}}>_i$, whereas the bottom-ranked cluster had the least favorable $<\Delta\Delta E^{\text{elec}}>_i$. Among the top 10% $<\Delta\Delta E^{\text{elec}}>_i$-ranked surface patches, the three patches with the largest $<C>_i$ values were selected and the constituent solvent-accessible residues were predicted to bind DNA.

Performance measures

To evaluate the performance of DR_bind, the numbers of correctly predicted binding residues (TP) and non-binding residues (TN), as well as the numbers of incorrectly predicted binding residues (FP) and non-binding residues (FN) were computed and used to determine:

$$\text{sensitivity} = \text{TP}/(\text{TP}+\text{FN}) \quad (4)$$

$$\text{specificity} = \text{TN}/(\text{FP}+\text{TN}), \quad (5)$$

$$\text{precision} = \text{TP}/(\text{TP}+\text{FP}) \quad (6)$$

$$\text{accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{FP}+\text{TN}+\text{FN}) \quad (7)$$

$$\text{Matthew's correlation coefficient or MCC} = (\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})/ [(\text{TP}+\text{FP})\,(\text{TP}+\text{FN})\,(\text{TN}+\text{FP})\,(\text{TN}+\text{FN})]^{1/2} \quad (8)$$

DR_bind web server

Input

On the DR_bind web page http://dnasite.limlab.ibms.sinica.edu.tw/, users are given two options: For option A, users upload their own file in PDB format and the evolutionary data for their protein in ConSurf format or ask DR_bind to retrieve the evolutionary data from ConSurf. For option B, users enter the PDB code and chain identifier; if the conservation profile for the submitted protein structure has not been pre-calculated in the ConSurf-DB database (32), DR_bind will attempt to generate the ConSurf data automatically from the ConSurf server (41–43). If no ConSurf data can be generated, DR_bind will continue to predict DNA-binding residues based only on the protein 3D structure and inform the user of the missing ConSurf data on the Results page. For multiple submissions, we have provided a simple form that allows for nine PDB codes with chain identifiers to be defined. After users click on the 'submit' button, the input data is checked for consistency: Residues in the PDB file that do not correspond to the standard 20 amino acid are removed, as well as multiple alternative residue positions. If the input data pass these tests, then the prediction process is started and the user is taken to a web page where the results for the job(s) and their status on the DR-bind server can be monitored.

Output

When the DR_bind server has finished the prediction, the results page is updated with the predicted binding

Figure 1. An example of the Results page from DR_bind.

residues. If the user had provided an e-mail address, the web server will send an e-mail to let the user know that the prediction has been completed with a link to the results web page. Users can then access the results page to see the generated prediction. As shown in Figure 1, the results page is split into three sections: the first section has links to downloadable files of (i) the original PDB and ConSurf files, (ii) the 'cleaned' PDB file used by DR_bind, (iii) a PyMOL script for highlighting the predicted DNA-binding residues and (iv) a text file of these residues. The second section lists the predicted DNA-binding residues. The third section is an interactive embedded 3D representation of the protein with the entire backbone in ribbon format with the predicted interaction residues depicted in stick format in red. This 3D representation is created using Jmol (http://www.jmol.org/) and can be rotated and zoomed in/out on the results page itself.

DR_bind currently runs on an Apple Mac Mini quad-core i7 server and the time taken to yield a prediction depends on the number of residues in the PDB chain. A prediction takes ~5 min for 50 residues, ~1.5 h for 200 residues, ~4.5 h for 350 residues and ~10 h for 450 residues. To handle simultaneous requests, the Torque batch processing software is used to queue jobs. Help pages with instructions on how to use the server are available at http://dnasite.limlab.ibms.sinica.edu.tw/examples/help.html.

RESULTS AND DISCUSSION

Performance and limitations of DR_bind

In our previous works (27), we presented a method for predicting DNA-binding sites based on electrostatics,

conservation and geometry given the respective protein structure and tested it on a set of 56 structurally non-homologous proteins with DNA-bound structures, as well as a smaller subset of 23 proteins with both DNA-bound and free structures. Based on the DNA-free and DNA-bound protein structures, 83 and 86% of the DNA-binding proteins have statistically significant DNA-binding sites, respectively. Thus, the method was found not to be very sensitive to protein conformational changes upon DNA binding (27,44). However, like all structure-based prediction methods, it cannot predict binding residues in regions that are disordered in the free protein structure. Another limitation of the method is that the predicted residues may be involved in binding non-DNA ligands such as RNA, protein, small molecules or metal ions rather than DNA (27,44).

In this work, we have implemented our DNA-binding residue prediction method as a free web server called DR_bind, which requires as input, the protein 3D structure and yields as output, experimentally testable residues that are predicted to bind DNA. As more DNA-binding protein structures have been solved since validation of our method (27), and some of these may correspond to novel folds, DR_bind was further tested using our updated set of 83 DNA-bound and 55 bound–unbound non-homologous protein structures, as well as the protein–DNA benchmark version 1.2 containing 47 bound–unbound structures (28). DR_bind yielded 47% precision, 35% sensitivity, 96% specificity, 90% accuracy and 35% mcc in predicting DNA-binding residues using our bound data set, and slightly lower precision (43%) and mcc (33%) values using our free data set (Table 1), even though the RMSD of the DNA-free structure from the respective DNA-bound structure may be as large as 33 Å (Supplementary Table S1). Similar trends were found for the benchmark data set: DR_bind yielded 56% precision, 40% sensitivity, 95% specificity, 87% accuracy and 40% mcc using the DNA-bound structures and lower precision (49%) and mcc (35%) values using the corresponding free structures (Table 1). The sensitivity values are low, as DR_bind predicts the most likely DNA-binding residues, rather than all DNA-binding residues at the protein–DNA interface.

To assess the reliability of the performance values in Table 1, we randomly chose 40 of the 83 DNA-bound structures and 25 of the 55 DNA-free protein structures and computed the various performance measures; this procedure was repeated 1000 times in order to obtain the distribution of each performance measure. Figure 2a and b illustrates the percent frequency of the DR_bind's precision values (solid lines) for the bound and free data sets, respectively. The lower limits of precision, sensitivity, specificity, accuracy and mcc in predicting DNA-binding residues using DR_bind for the bound/free data sets are 0.38/0.31, 0.29/0.26, 0.94/0.93, 0.87/0.86 and 0.29/0.24, whereas the corresponding upper limits are 0.56/0.55, 0.44/0.49, 0.97/0.97, 0.91/0.92 and 0.43/0.44. Notably,

Table 1. Comparison of the performance measures of DR_Bind using our nonredundant data set of 83 DNA-bound and 55 DNA-free protein structures and the protein–DNA benchmark version 1.2 containing 47 DNA-bound and free protein structures

Data set	I (bound)	II (free)	III (bound)	III (free)
No. of structures	83	55	47	47
TP	728	419	468	417
FP	831	566	371	429
TN	18 128	11 596	6486	6435
FN	1,362	792	702	693
Precision	0.47	0.43	0.56	0.49
Sensitivity	0.35	0.35	0.40	0.38
Specificity	0.96	0.95	0.95	0.94
Accuracy	0.90	0.90	0.87	0.86
mcc	0.35	0.33	0.40	0.35

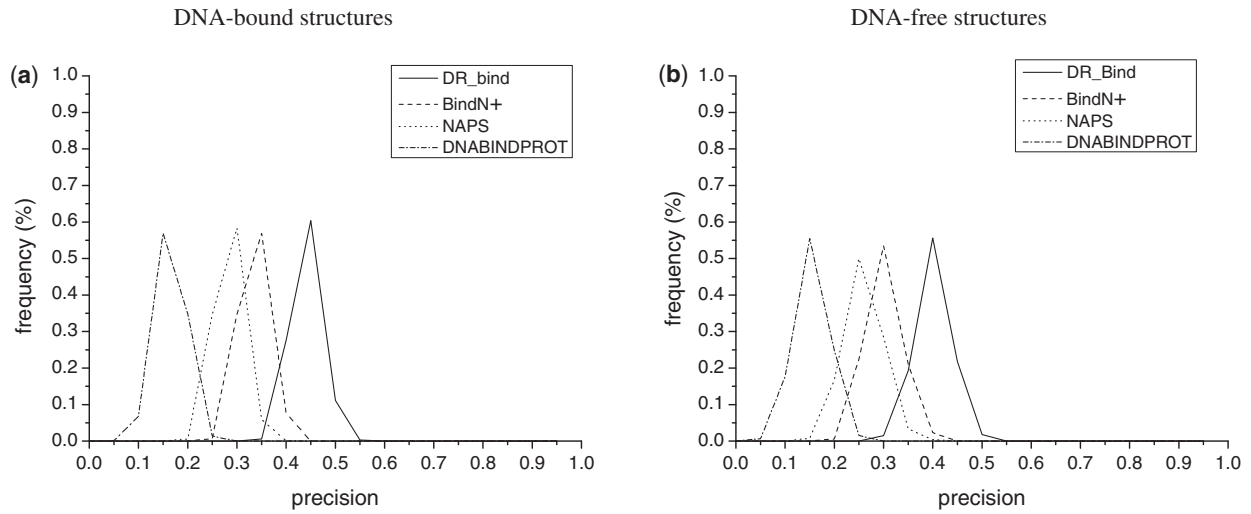

Figure 2. The percent frequency of a precision value derived from 1000 random choices of (**a**) 40 DNA-bound structures from Data set I and (**b**) 25 DNA-free protein structures from Dataset II. The solid, dashed, dotted and dashed–dotted curves correspond to precision values obtained using DR_bind, BindN+, NAPS and DNABINDPROT, respectively.

these limits encompass the precision, sensitivity, specificity, accuracy and mcc values obtained using the 47 bound–unbound structures from the benchmark data set.

Comparisons with other servers that predict DNA-binding residues

Using our bound and free data sets, the performance of DR_bind was compared with that of three recent web servers, BINDN+ (http://bioinfo.ggc.org/bindn+/), NAPS (http://proteomics.bioengr.uic.edu/NAPS) and DNABINDPROT (http://www.prc.boun.edu.tr/appserv/prc/dnabindprot/). BINDN+ (12) uses support vector machines with three biochemical features (hydrophobicity, side chain pKa and mass of an amino acid residue) incorporating evolutionary information and position-specific scoring matrix (PSSM). Instead of support vector machines, NAPS (15) employs ensemble classifiers based on C4.5, bootstrap aggregation and a cost-sensitive learning algorithm with residue charge and PSSM. Whereas BINDN+ and NAPS are sequence-based methods, DNABINDPROT (23) is a structure-based method that identifies high-frequency fluctuating conserved residues and ranks them according to their DNA-binding propensity. These web servers were chosen for comparison with DR_bind because they had been tested using published data sets and had been shown to outperform previous methods/web servers: Using the PDNA-62 data set, the average of sensitivity and specificity obtained by BINDN+ (78.3%) and NAPS (78.5%) were similar (12,15) and higher than that obtained by DP-Bind (76.5%) or DBS-PSSM (67.1%). Using a set of 36 DNA-binding proteins with both free and DNA-bound structures and conservation scores, the precision obtained by DNABINDPROT using a fast threshold of 0.1, conservation threshold of 5, and neighboring two residues (45.3%) was higher than that obtained by DBD-HUNTER (44.5%), DISPLAR (40%) and DP-Bind (33.0%) (23).

Using our bound and free data sets, the performance results of all four servers are summarized in Table 2. Since DR_Bind does not aim to predict all residues at the protein–DNA interface, its sensitivity (35%) is lower than that of BINDN+ (45–48%), which has almost twice the number of predictions (i.e. TP + FP). Rather than knowing all residues that comprise the protein–DNA interface, most biologists would be interested in testing if the predicted residues do indeed bind DNA and therefore, a method's precision, which reflects the fraction of predicted residues that are correct. Compared with the other methods, DR_Bind yields a ≥10% higher precision for both data sets. To assess if the difference in precision using DR_Bind and the other three methods is statistically significant, we randomly chose 40 and 25 protein structures from the bound and free data sets, respectively, and computed the precision obtained by each of the four servers; this was repeated 1000 times. The precision values obtained by DR_bind using the DNA-bound (0.38–0.56) and DNA-free structures (0.31–0.55) are generally higher than those obtained by the other three methods, as shown in Figure 2. This is also shown by the paired *t*-test, which was used to test the null hypothesis that DR_Bind does 'not' yield higher precision than the other three methods. The resulting $P < 0.00001$ for both bound and free data sets rejected the null hypothesis (Supplementary Table S3). Hence, an experimentalist would likely find more residues predicted by DR_bind to bind DNA compared with those predicted by sequence-based methods, thus saving time and costs.

Compared with sequence-based methods to predict DNA-binding residues, the structure-based DR_bind approach incorporates structural information (that is, electrostatics and geometry) of the query protein. Therefore, it would be expected to perform much better than sequence-based methods when evolutionary information for a query protein is not available. To show the importance of additional structural information, we tested the structure- and sequence-based methods on a set of 15 non-redundant DNA-bound protein structures with no or unreliable ConSurf conservation profiles. Note that DNABINDPROT could not be applied to this set of 'unique' DNA-binding proteins because it does not yield predictions for proteins without ConSurf-DB conservation data. The performance results of DR_bind, BINDN+ and NAPS in Table 3 show that the difference in performance between DR_bind and the two sequence-based methods become more apparent for proteins without conservation data: the precision of DR_bind (47%) is nearly twice that of BINDN+ (27%) and NAPS (23%). Thus, for DNA-binding proteins with no or insufficient homologs, DR_bind could provide a significantly higher fraction of correctly predicted DNA-binding residues than sequence-based methods.

Table 2. Comparison of the performance measures of DR_Bind, BindN+, NAPS and DNABINDPROT using the same data set of 83 DNA bound[a] or 55 DNA-free protein structures[b,c]

Server	DR_Bind	BindN+	NAPS	DNABINDPROT
TP	728 (419)	1013 (542)	328 (180)	244 (169)
FP	831 (566)	1798 (1129)	733 (459)	1040 (772)
TN	18 128 (11 596)	17 161 (11 033)	18 226 (11 703)	17 919 (11 390)
FN	1362 (792)	1077 (669)	1762 (1031)	1846 (1042)
Precision	0.47 (0.43)	0.36 (0.32)	0.31 (0.28)	0.19 (0.18)
Sensitivity	0.35 (0.35)	0.48 (0.45)	0.16 (0.15)	0.12 (0.14)
Specificity	0.96 (0.95)	0.91 (0.91)	0.96 (0.96)	0.95 (0.94)
Accuracy	0.90 (0.90)	0.86 (0.87)	0.88 (0.89)	0.86 (0.86)
mcc	0.35 (0.33)	0.34 (0.31)	0.16 (0.15)	0.08 (0.09)

[a]The PDB entries are listed in Supplementary Table S1; the total number of residues in the data set is 21 049, out of which 2090 residues are DNA-binding (=TP+FN) and 18 959 residues are non-DNA-binding (=FP+TN).
[b]Performance measures based on the DNA-free protein structures are in the parentheses.
[c]The PDB entries are listed in Supplementary Table S1; the total number of residues in the dataset is 13 373, out of which 1211 residues are DNA-binding (=TP+FN) and 12 162 residues are non-DNA-binding (=FP+TN).

Table 3. Comparison of the performance measures of DR_Bind, BindN+ and NAPS using the same data set of 15 DNA-bound protein structures with no or insufficient close homologs[a]

Server	DR_Bind	BindN+	NAPS
TP	110	230	34
FP	122	618	115
TN	2585	2089	2592
FN	292	172	368
Precision	0.47	0.27	0.23
Sensitivity	0.27	0.57	0.08
Specificity	0.95	0.77	0.96
Accuracy	0.87	0.75	0.84
mcc	0.29	0.26	0.07

[a]The PDB entries are listed in Supplementary Table S2; the total number of residues in the data set is 3109, out of which 402 residues are DNA-binding (=TP+FN) and 2707 residues are non DNA-binding (=FP+TN).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3.

ACKNOWLEDGEMENTS

We thank Karen Sargsyan for helpful discussion.

FUNDING

Academia Sinica and the National Science Council, Taiwan. Funding for open access charge: National Science Council, Taiwan [NSC 95-2113-M-001-038-MY5] (to C.L.).

Conflict of interest statement. None declared.

REFERENCES

1. Strong,M.J., Volkening,K., Hammond,R., Yang,W., Strong,W., Leystra-Lantz,C. and Shoesmith,C. (2007) TDP43 is a human low molecular weight neurofilament (*h*NFL) mRNA-binding protein. *Mol. Cell. Neurosci.*, **35**, 320–327.
2. Pavletich,N.P., Chambers,K.A. and Pabo,C.O. (1993) The DNA-binding domain of p53 contains the four conserved regions and the major mutation hot spots. *Genes Dev.*, **7**, 2556–2564.
3. Ahmad,S., Gromiha,M.M. and Sarai,A. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
4. Keil,M., Exner,T.E. and Brickmann,J. (2004) Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network. *J. Comput. Chem.*, **25**, 779–789.
5. Ahmad,S. and Sarai,A. (2005) PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics*, **6**, 33.
6. Yan,C., Terribilini,M., Wu,F., Jernigan,R., Dobbs,D. and Honavar,V. (2006) Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, **7**, 262.
7. Kuznetsov,I.B., Gou,Z., Li,R. and Hwang,S. (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins*, **64**, 19–27.
8. Hwang,S., Gou,Z. and Kuznetsov,I.B. (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics*, **23**, 634.
9. Ofran,Y., Mysore,V. and Rost,B. (2007) Prediction of DNA-binding residues from sequence. *Bioinformatics*, **23**, i347–i353.
10. Chu,W., Huang,Y., Huang,C., Cheng,Y., Huang,C. and Oyang,Y. (2009) ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors. *Nucleic Acids Res.*, **37**, W396–W401.
11. Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, **34**, W243–W248.
12. Wang,L., Huang,C., Yang,M. and Yang,J. (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.*, **4**, S3.
13. Wu,J., Liu,H., Duan,X., Ding,Y., Wu,H., Bai,Y. and Sun,X. (2009) Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics*, **25**, 30–35.
14. Wang,L., Yang,M.Q. and Yang,J.Y. (2009) Prediction of DNA-binding residues from protein sequence information using random forest. *BMC Genomics*, **10**, S1.
15. Carson,M.B., Langlois,R. and Lu,H. (2010) NAPS: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Res.*, **38**, W431–W435.
16. Si,J., Zhang,Z., Lin,B., Schroeder,M. and Huang,B. (2011) MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. *BMC Syst. Biol.*, **5**, S7.
17. Jones,S., Shanahan,H.P., Berman,H.M. and Thornton,J.M. (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.*, **31**, 7189–7198.
18. Stawiski,E.W., Gregoret,L.M. and Mandel-Gutfreund,Y. (2003) Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol.*, **326**, 1065–1079.
19. Tsuchiya,Y., Kinoshita,K. and Nakamura,H. (2004) Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins*, **55**, 885–894.
20. Shanahan,H.P., Garcia,M.A., Jones,S. and Thornton,J.M. (2004) Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Res.*, **32**, 4732–4741.
21. Ferrer-Costa,C., Shanahan,H.P., Jones,S. and Thornton,J.M. (2005) HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif. *Bioinformatics*, **21**, 3679–3680.
22. Wu,C.Y., Chen,Y.C. and Lim,C. (2010) A structural-alphabet-based strategy for finding structural motifs across protein families. *Nucleic Acids Res.*, **38**, e150.
23. Ozbek,P., Soner,S., Erman,B. and Haliloglu,T. (2010) DNABINDPROT: fluctuation-based predictor of DNA-binding residues within a network of interacting residues. *Nucleic Acids Res.*, **38**, W417–W423.
24. Tsuchiya,Y., Kinoshita,K. and Nakamura,H. (2005) PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces. *Bioinformatics*, **21**, 1721–1723.
25. Tjong,H. and Zhou,H.X. (2007) DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res.*, **35**, 1465.
26. Gao,M. and Skolnick,J. (2008) DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. *Nucleic Acids Res.*, **36**, 3978–3992.
27. Chen,Y.C., Wu,C.Y. and Lim,C. (2007) Predicting DNA-binding sites on proteins from electrostatic stabilization upon mutation to Asp/Glu and evolutionary conservation. *Proteins*, **67**, 671–680.
28. van Dijk,M. and Bonvin,A.M.J.J. (2008) A protein–DNA docking benchmark. *Nucleic Acids Res.*, **36**, e88.
29. Chen,Y.C. and Lim,C. (2008) Common physical basis of macromolecule-binding sites in proteins. *Nucleic Acids Res.*, **36**, 7078–7087.
30. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Iype,L., Jain,S., Fagan,P., Marvin,J. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D*, **58**, 899–907.
31. Pearl,F., Todd,A., Sillitoe,I., Dibley,M., Redfern,O., Lewis,T., Bennett,C., Marsden,R., Grant,A., Lee,D. *et al.* (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.*, **33**, D247–D251.

32. Goldenberg,O., Erez,E., Nimrod,G. and Ben-Tal,N. (2009) The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.*, **37**, D323–D327.
33. Taylor,W.R. and Orengo,C.A. (1989) Protein structure alignment. *J. Mol. Biol.*, **208**, 1–22.
34. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
35. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
36. Miller,S., Janin,J., Lesk,A.M. and Chothia,C. (1987) Interior and surface of monomeric proteins. *J. Mol. Biol.*, **196**, 641–656.
37. Koradi,R., Billeter,M. and Wuthrich,K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.*, **14**, 51–55.
38. Canutescu,A.A., Shelenkov,A.A. and Dunbrack,R.L. Jr (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.*, **12**, 2001–2014.
39. Case,D.A., Cheatham,T.E. III, Darden,T., Gohlke,H., Luo,R., Merz,K.M. Jr, Onufriev,A., Simmerling,C., Wang,B. and Woods,R.J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688.
40. Duan,Y., Wu,C., Chowdhury,S., Lee,M.C., Xiong,G., Zhang,W., Yang,R., Cieplak,P., Luo,R., Lee,T. *et al.* (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, **24**, 1999–2012.
41. Ashkenazy,H., Erez,E., Martz,E., Pupko,T. and Ben-Tal,N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–W533.
42. Landau,M., Mayrose,I., Rosenberg,Y., Glaser,F., Martz,E., Pupko,T. and Ben-Tal,N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, 299–302.
43. Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor-Shental,D., Martz,E. and Ben-Tal,N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
44. Chen,Y.C. and Lim,C. (2008) Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry. *Nucleic Acids Res.*, **36**, e29.